

HPC@Jülich as opportunistic resource for CMS

Integration prototyping

T. Madlener, C. Wissing, M. Giffels

08.03.2022




Overview / Introduction

- Final goal: Integrate FZJ resources into the CMS pool
 - Ideally transparently, e.g. via T1@KIT
- Current goal: Have a running prototype on the ITB pool
- HPC resources usually quite different from the “usual grid resources”
 - Networking much more restricted
 - Available software differs, e.g. no CVMFS by default
- Situation on JURECA
 - CVMFS setup to get CMS software
 - Network (proxy) setup
 - Launching manual glideins
- Cobald / Tardis setup
- Issues still to be resolved


JURECA @ FZJ

- Compute nodes with AMD EPYC 7742 (2x64 @ 2.25 GHz) with either 512 or 1024 GB of RAM + potentially Nvidia A100 GPUs
 - [JURECA hardware configuration](#)
 - Running on CPU only at the moment
 - Seems to be the resource with the least network restrictions
- Running *Rocky Linux 8.5* (CentOS8)
 - Singularity 3.8.5 (without user namespaces)
 - Slurm batch system
- Associated to the *csInpp* project (lead by Stefan Krieg)
 - Have 200k core hours more or less for ourselves on JURECA CPU nodes
 - Try to get a basic workflow to work and do some first tests

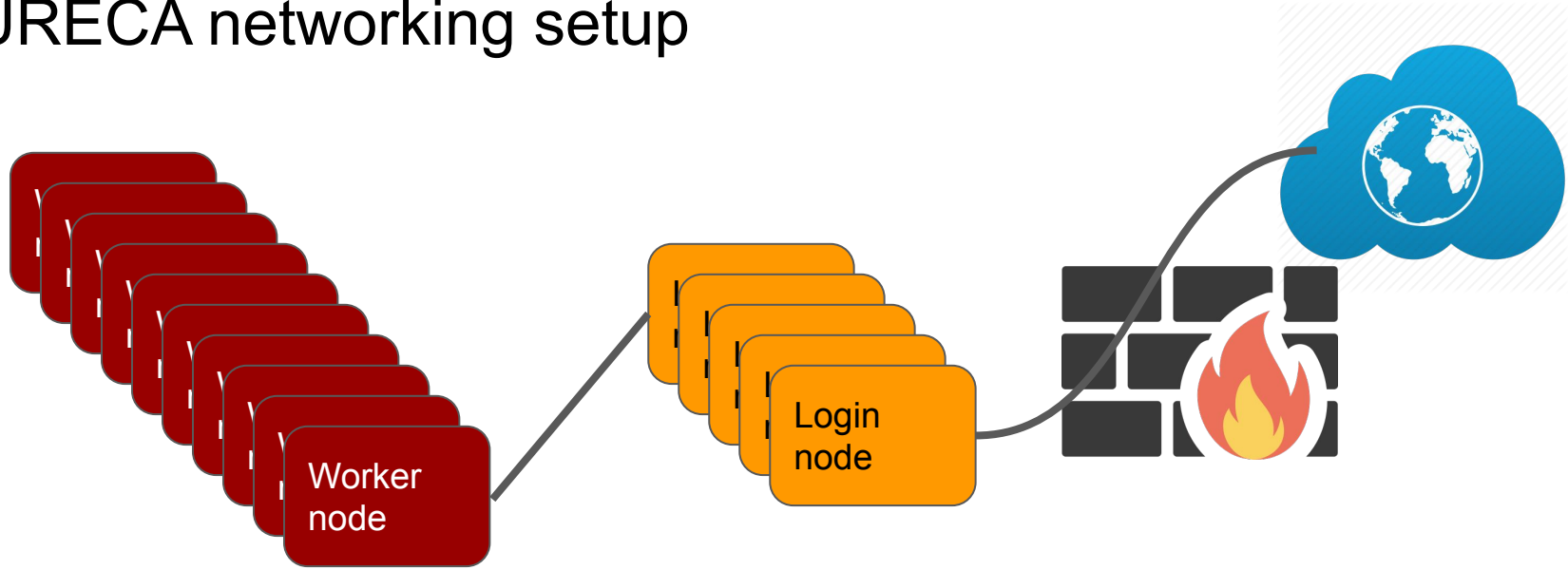
Frontier squid setup

- Possible to run a squid proxy on the login node directly and connect to it from worker nodes 
 - For testing purposes only
 - Needs to be moved to a dedicated node or VM for production, but don't expect too many problems with that
- Also using squid proxy as CVMFS http_proxy
 - Makes CVMFS setup on worker nodes slightly easier

CVMFS and container setup

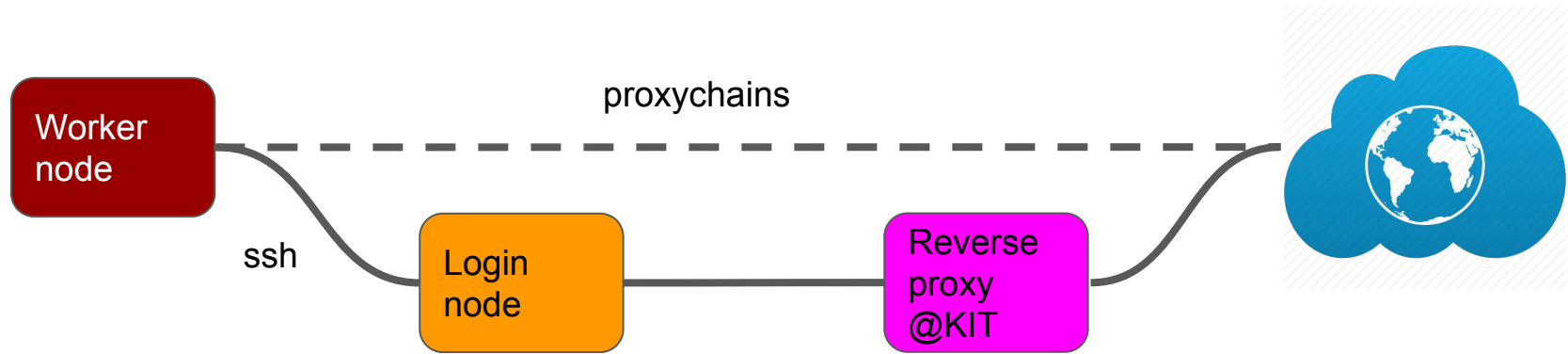
- CVMFS is not mounted on JURECA
- Use [cvmfsexec](#) to mount it locally and then bind-mount it to /cvmfs/ in a singularity container 
- Image is cmssw/cc7:amd64 with some additional packages for our network setup (and debugging), [Dockerfile](#)
- Singularity on JURECA has user namespaces disabled
 - Not possible to nest containers
 - Not possible to use images from unpacked.cern.ch
 - Need dedicated glidein that allows to run CMS workflows without starting their own singularity container

JURECA networking setup



- Worker nodes can only connect to Login nodes
- Login nodes have outbound connections
 - But only on ports 80 (http) and 443 (https)
 - We need other ports as well

Current proxy setup using proxychains



- ssh tunnel from worker node to login node
- Use proxychains to route all traffic through ssh tunnel and port 80 of login node to
- Reverse proxy @KIT running dante2 proxy on a VM with less strict firewall

[proxychains setup](#)

Putting everything into a batch job

[script](#)

- Mount cvmfs repositories via cvmfsexec
- Setup ssh tunnel from worker to login node
- Launch singularity container and bind-mount cvmfs repositories to /cvmfs


“Setup” on worker node

- Enter the now running container

- Pre-load proxychains library
- Launch (manual) glidein
- Enjoy

Work that is done inside the container that is launched in the setup

Cobald / Tardis setup

- In the end do not want to have to launch glidein batch jobs manually but delegate that to Cobal / Tardis
- Running on a VM @FZJ
 - Access to ITB schedd (vocms068)
 - Access to jureca login node
- Submission of batch jobs and monitoring of running glideins on JURECA via ssh
- [Small hack](#) of glidein startup scripts necessary to inject the *TardisDroneUuid*
- Successfully tested on a rather small scale 
 - Submitting enough jobs to fill 1 node completely (64 jobs)

Workflow for testing

- Using a [MC workflow](#) for testing
 - Needs no additional input
 - Can be tuned quite easily to different runtimes
 - Also used, e.g. by CINECA for initial testing and debugging
- Submitted to ITB pool via crab
- Targeting FZJ glideins via *GLIDEIN_Entry_Name* ClassAd and extraJDL
- Storage of outputs at T2_DE_DESY

Open issues

- Issues to be addressed for successful prototyping
 - Payloads sent to glideins lose connection to shadow on schedd
 - FZJ firewall seems to kill (idle?) outgoing connections (on port 80?)
 - Jobs run successfully, including stageout
 - Payloads cannot report status back to schedd and are “lost”
 - Scale tests of the whole setup with more job pressure (roughly 70k core hours still available)
- Issues to be addressed on a longer term scale (i.e. before production)
 - Test setup is using a personal grid certificate for glidein authentication with the ITB pool
 - Properly upstream the TardisDroneUuid injection code
 - Check with FZJ admins if some parts of the setup could be simplified with a bit of their help
 - Need a new glidein for being able to use *customisable pilots* that would potentially allow for an easier selection of jobs to run on FZJ
 - Migrate squid proxy to its own node or VM

Summary & Outlook

- Currently working on getting a prototype workflow to run on JURECA @FZJ
- Working setup for CVMFS access
- Need to work around some network restrictions
 - Proxychains based solution almost working
 - Need to figure out a way to avoid firewall timeouts
- Can run CMS MC generation workflow, including stageout
 - Reporting success back to pool not working due to network problems
- First tests with using Cobald / Tardis for automatic submission of glideins
- (Small) scale tests as soon as we have stable network setup
- [repository with the necessary setup code](#)